KEY_Practice24_Basic_Stats_IV_Significance

December 10, 2021

1 Practice: Statistical Significance

Let's continue to work with the diabetes dataset to apply a t-test to real world data.

```
[1]: # Import pandas, so that we can import the diabetes dataset and work with the␣
     ↪data frame version of this data
     import pandas as pd
```

```
[2]: # Set the path
     path = 'https://raw.githubusercontent.com/GWC-DCMB/curriculum-notebooks/master/'
     # This is where the file is located
     filename = path + 'SampleData/diabetes.csv'
```

```
[3]: # Load the diabetes dataset into a DataFrame
     diabetes_df = pd.read_csv(filename)
     diabetes_df
```

```
[3]:      AGE  SEX   BMI     MAP   TC    LDL   HDL   TCH     LTG  GLU    Y
     0     59    2  32.1  101.00  157   93.2  38.0  4.00  4.8598   87  151
     1     48    1  21.6   87.00  183  103.2  70.0  3.00  3.8918   69   75
     2     72    2  30.5   93.00  156   93.6  41.0  4.00  4.6728   85  141
     3     24    1  25.3   84.00  198  131.4  40.0  5.00  4.8903   89  206
     4     50    1  23.0  101.00  192  125.4  52.0  4.00  4.2905   80  135
     ..   …   …    …      …    …     …     …     …      …    …   …
     437   60    2  28.2  112.00  185  113.8  42.0  4.00  4.9836   93  178
     438   47    2  24.9   75.00  225  166.0  42.0  5.00  4.4427  102  104
     439   60    2  24.9   99.67  162  106.6  43.0  3.77  4.1271   95  132
     440   36    1  30.0   95.00  201  125.2  42.0  4.79  5.1299   85  220
     441   36    1  19.6   71.00  250  133.2  97.0  3.00  4.5951   92   57

     [442 rows x 11 columns]
```

We are interested in understanding whether there are differences in LDL levels (the "bad" cholesterol) by sex, i.e. are LDL levels different for males vs. females?

1. Formulate the null hypothesis and the alternative hypothesis. - **Null hypothesis**: There is NO difference in LDL levels between male and female. - **Alternative hypothesis**: There is a difference in LDL levels by sex.

[4]:

```python
# Import numpy
import numpy as np
```

Males are indicated by “1” for the variable “SEX”, while females are indicated by “2”.

[5]:

```python
# Define a vector of the LDL levels for males and name it ldl_male
diabetes_male = diabetes_df.query('SEX == 1')
ldl_male = diabetes_male['LDL']

# Define a vector of the LDL levels for females and name it ldl_female
diabetes_female = diabetes_df.query('SEX == 2')
ldl_female = diabetes_female['LDL']
```

2. Identify and compute a test statistic that can be used to reject or fail to reject the null hypothesis. - As we are working with two independent samples, we will use the two-sample t-test and use the t-statistic.

3. Compute the test statistic and p-value.

[6]:

```python
# Import stats methods to help calculate the t-statistic and p-value
from scipy import stats
```

[7]:

```python
# Run a Student's t-test
t_statistic, p_value = stats.ttest_ind(ldl_male, ldl_female)

# Print out the test statistic and p-value
print("t-statistic = " + str(t_statistic))
print("p-value = " + str(p_value))
```

```
t-statistic = -3.022893334345971
p-value = 0.0026499873735660695
```

4. Compare the p-value to an acceptable significance value, α and compare the test statistic to acceptable critical value(s). If p-value $\leq \alpha$ and the test-statistic $\geq$ +critical value or test-statistic $\leq$ -critical value, that the observed effect is statistically significant, the null hypothesis is rejected, and the alternative hypothesis is valid.** - p-value $= 0.0026 < 0.05$, so we reject the null hypothesis. - t-statistic $= -3.02 < -1.96$, so this reaffirms that we reject the null hypothesis. - Interpretation: There is a significant difference in LDL levels between males and females.

Congratulations on completing the lesson and practice!

It’s a lot of information, but you learned powerful tools to be on your way to answer your own research questions by analyzing real world data!

Challenge: Using the code you wrote above as a template, can you run a t-test comparing LDL Cholesterol for people 50 & older vs. people under 50?

[8]:

```python
# Define a vector of the LDL levels for people 50 or older
diabetes_over50 = diabetes_df.query('AGE >= 50')
```

```python
ldl_over50 = diabetes_over50['LDL']

# Define a vector of the LDL levels for females and name it ldl_female
diabetes_under50 = diabetes_df.query('AGE < 50')
ldl_under50 = diabetes_under50['LDL']

# Run a Student's t-test
t_statistic, p_value = stats.ttest_ind(ldl_over50, ldl_under50)

# Print out the test statistic and p-value
print("t-statistic = " + str(t_statistic))
print("p-value = " + str(p_value))
```

```
t-statistic = 3.185760417933572
p-value = 0.001546465356577734
```